HIGH ACCURACY AND HIGH PRECISION HLA TYPING USING ILLUMINA READS:

TYPING ALGORITHM AND PERSPECTIVES ON NOVEL ALLELE DISCOVERY

Authors: György Horváth, Krisztina Rigó, Tim Hague, Attila Bérces, Szilveszter Juhos
Omixon Biocomputing Ltd., Budapest, Hungary

Introduction

HLA genes are both polymorphic and homologous - besides the differences in their sequence there can be regions that are similar not only for different alleles but also for different genes. Therefore the classic de-novo approach used in genomics to reconstruct the consensus sequences in many cases doesn't work for HLA genes. To make NGS based typing even more cumbersome, the whole genomic reference sequence is not available for all alleles, for many only the coding parts (CDS) or the most polymorphic exons are present in the IMGT/HLA database. Our consensus based HLA typing method aims to deliver phased consensus sequences and associated alleles, supported by the most reads of the provided sample data.

SAMPLE DATA

Short reads

Subsampling and filtering

Filtering and preprocessing sample reads increases the average data quality and improves the processing performance. Homogeneous subsampling of the input data ensures that the resulting read selection is reflecting the original distribution regarding the genes and the allele balance. Quality based read trimming with a bidirectional sliding window method is applied, also reads which are too short get dropped.

Parallel mapping and alignment merge

Mapping short reads on the IMGT/HLA database of known alleles and using this initial layout information of the reads as the basis of further steps. The mapping is executed in parallel for all alleles and redundant mappings are preserved. The resulting alignments get partly merged to reduce resource consumption. We take use of this alignment step only for layouting/positioning reads - no allele-level information is considered later, as we are trying to be independent from database issues as much as possible. We only assume that the sample data which we are analyzing can be reproduced by freely mixing a set of known alleles in the database plus applying a set of variations if needed.

Read layout

First round variant call and branching

Running an initial variant call on each region. As a result in some cases we get an unambiguous resolution, in some cases there are positions which should be treated as gaps due to low coverage and in some cases there are positions where multiple resolution branches could be identified. To get phased consensus fragments we have to permute these branches at these positions and check for each permutation if it has enough supporting reads and refined variant call can be executed. As a result of this step we know which read groups should be inspected for consensus fragment generation.

Read groups

Second round variant call and sequence generation

Once read groups with coherent information are established, a refined variant call can be executed to verify the consistency at each position and to generate initial consensus fragments only with sufficient, contradiction free read support. The most important properties of these initial consensus fragments are: random noise is filtered out, the supporting read count at each base ensures that only relevant signal is kept and the consensus sequence is phased therefore it can be used for proper phase resolution later for larger regions.

Fragment merge and reduction

The phased initial fragment set created previously gets shrinked iteratively while we maintain the properties of the fragments but reduce the number of them. At each reduction step we take care of not losing information. Note that contamination and sequencing artifacts are still part of the processed fragment set. Accurate tracking of the reads forming the fragments is important to retrieve read support information later while calculating variant frequencies.

Initial fragments

Filtering contamination and PCR crossover artifacts, ploidy resolution

For ploidy resolution - where we cluster the fragments by originating chromosome - we need to sort out those fragments which don't belong to the biological sample. The most important sources of fragments with incompatible content are: contamination and PCR crossover artifacts.

Contamination fragments are discovered by measuring the related variant frequencies and comparing to variants identified in other fragments.

Fragments built from cross-over reads can be detected by comparing them to other fragments on both chromosomes in the same region sharing some variations.

Heterozygous positions are the key to assign fragments to chromosomes. The assignment method is based on the fragment pair test which has the following simple criteria:

- fragments must be overlapping according to their pairwise alignment,
- fragments must not be matching, at least one variation should exist in the overlapping region.

Based on the pair test we can create a bipartite graph: nodes are consensus fragments and edges are marking the successful pair tests. This graph provides the basis of the consensus component formation.

Merged fragments

Fragment layout resolution

This is the step where we start to discover the layout of the fragments. We organize fragments into pileups which provides the base for further steps. Each pileup contains a set of fragments aligned together, ordered by start and end positions. Both multiple sequence alignment for the entire pileup and refined pairwise alignments for overlapping elements are calculated. A single consensus generation region (e.g. intron 3) might consist of multiple pileups if the pileups are separated (e.g. due to amplicon boundaries) or just bound together weakly (usually due to a gap in the coverage).

Fragment pileup

Fragment merge, phasing resolution and layout finalization

Once we have the fragment-chromosome assignments we can start building the final consensus layout. Fragments within the same region on the same chromosome can now be concatenated according to the pileup layout. Homozygous fragments are added to all components in the same region. Phase resolution also takes place here. The final layout creation handles alternative splicing cases too.

Consensus components

1 2 ?

Searching matching alleles

As the final step of the typing process we search for alleles matching the generated consensus resolution. We are reporting allele pairs where phasing and complementary usage of the chromosome consensus sequences are considered to ensure that all available consensus information is matching the involved reference alleles. The result allele pairs are ordered by mismatch count. In some cases there are multiple alternative results with the same mismatch count due to phasing or splicing ambiguity.

Consensus sequences

C G A C T G A
C C G A G T A

RESULT BROWSER

Results and conclusion

Using consensus based typing for samples targeting genomic HLA-A, -B, -C, -DRB1 and -DQB1 genes we managed to get 100% concordance compared to Sanger SBT with very little ambiguity and usually 8 digits resolution (15 samples both pooled and unpooled).

For another test set (8 samples, pooled for HLA-A, -B, -C and DRB1) covering the genes partially we also managed to get 100% concordance but the ambiguities were higher.

For most of the samples with high ambiguity, consensus based typing revealed that the actual cause of ambiguity is the presence of novel alleles - most of the novelties are in intronic parts.

Corresponding author:

György Horváth (gyorgy.horvath@omixon.com)